Genomics 92 (2008) 414–418

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Analysis of human disease genes in the context of gene essentiality

Donghyun Park [e], Jungsun Park [c], Seung Gu Park [a], Taesung Park [c,d], Sun Shim Choi [a,b,*]

[a] Department of Molecular and Medical Biotechnology, Kangwon National University, Chunchon 200-701, Korea
[b] Institute of Bioscience and Biotechnology, Kangwon National University, Chunchon 200-701, Korea
[c] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea
[d] Department of Statistics, Seoul National University, Seoul 151-742, Korea
[e] Howard Huges Medical Institute, Department of Human Genetics, University of Chicago, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 24 March 2008
Accepted 7 August 2008
Available online 1 October 2008

Keywords:
Human disease gene
Gene essentiality
Disease mutation
Phenotype
Mouse
Nonsynonymous/synonymous rate ratio

ABSTRACT

The characteristics of human disease genes were investigated through a comparative analysis with mouse mutant phenotype data. Mouse orthologs with mutations that resulted in discernible phenotypes were separated from mutations with no phenotypic defect, listing '*phenotype*' and '*no phenotype*' genes. First, we showed that *phenotype* genes are more likely to be disease genes compared to *no phenotype* genes. *Phenotype* genes were further divided into '*embryonic lethal*', '*postnatal lethal*', and '*non-lethal phenotype*' groups. Interestingly, *embryonic lethal* genes, the most essential genes in mouse, were less likely to be disease genes than *postnatal lethal* genes. These findings indicate that some extremely essential genes are less likely to be disease genes, although human disease genes tend to display characteristics of essential genes. We also showed that, in *lethal* groups, non-disease genes tend to evolve slower than disease genes indicating a strong purifying selection on non-disease genes in this group. In addition, *phenotype* and *no phenotype* groups showed differing types of disease mutations. Disease genes in the *no phenotype* group displayed a higher frequency of regulatory mutations while those in the *phenotype* group had more frequent coding mutations, indicating that the types of disease mutations vary depending on gene essentiality. Furthermore, missense disease mutations in *no phenotype* genes were found to be more radical amino acid substitutions than those in *phenotype* genes.

Introduction

'Essential' and 'nonessential' are classic molecular genetic terminologies referring to the functional importance of a gene, specifically with regard to its fitness effect on an organism [1]. A gene that is absolutely required for survival, or a gene that strongly contributes to fitness and robust competitive growth is considered to be an essential gene [2]. On the contrary, a nonessential gene is dispensable for viability, and its inactivation yields viable and fertile individuals. Recently, several interesting findings have been made regarding the discernible characteristics of essential genes in model organisms. First, considering that strong purifying selection applied to essential genes would result in a lower evolutionary rate, essential genes have been shown to evolve more slowly than nonessential genes [1,3,4], although some conflicts have been reported [5,6]. Second, essential genes are likely to encode hub proteins in protein–protein interaction networks [7–10], suggesting that essential proteins have more interacting partners within networks than nonessential counterparts. Third, essential genes are more likely to be abundantly and ubiquitously expressed in cells and tissues [11,12] and have smaller-sized introns [13–15].

Human disease genes are considered to represent a subset of essential genes, given that heritable diseases generate discernible phenotypic symptoms caused by deleterious effects of disease mutations. To study this issue, various evolutionary features have been compared between heritable disease genes and non-disease genes [8,14,16,17]. Although most findings are generally consistent, conflicting results have been occasionally reported. Lopez-Bigas and Ouzounis [14] showed that human disease genes are more evolutionarily conserved than other genes, in agreement with the idea that genes with strong fitness effects should evolve more slowly than other genes. In contrast, Huang et al. [16] found that nonsynonymous/synonymous substitution rate ratios (K_a/K_s) in disease genes were similar to those in non-disease genes. Smith and Eyre-Walker [22] reported even higher K_a/K_s ratios in human disease genes than those in non-disease genes. In the latter study, human disease genes were shown to be expressed in a narrower range of tissues, contrary to the observed features of essential genes. Recently, it was reported that disease genes have a higher connectivity in networks than non-disease genes, which is a signature of essential genes. However, many human disease genes generate proteins located at the periphery of

* Corresponding author. Department of Molecular and Medical Biotechnology, Institute of Bioscience and Biotechnology, Kangwon National University, Chunchon 200-701, Republic of Korea. Fax: +82 33 241 6480.
E-mail address: schoi@kangwon.ac.kr (S.S. Choi).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.08.001

networks [8], indicating human disease genes are a mixture of both essential and nonessential genes. Some human genes essential in early development contribute to the high rate of first-trimester spontaneous abortions, which may account for as much as 20% of recognized pregnancies. Since those essential genes are very likely to be non-disease genes, human non-disease genes may also contain both essential and nonessential genes. Hence, Tu et al. [17] subtracted a set of defined essential genes (i.e., ubiquitously expressed housekeeping genes) from non-disease genes, concluding that disease genes have an intermediate essentiality between housekeeping genes and other human genes.

As it is impractical to experimentally estimate essentiality of human genes in the manners performed for model organisms, we utilized existing knowledge on essentiality of mouse orthologs as a strategy to estimate essentiality of human genes. Mouse mutant phenotype data are relatively extensive, and the evolutionary distance between human and mouse is relatively short. It has been widely accepted that homologous genes have similar functions if the two species are evolutionarily close, and that genes having essential functions in mice frequently have accordingly pivotal functions in humans. In the present study, we examined the likelihood of a gene being linked to human disease, in relation to the level of essentiality of the gene in mouse. In addition, by performing the analysis on the subgroups of disease genes with different levels of essentiality, we show that features of disease mutations on nonessential genes differ from those on essential genes.

Results

Gene classifications

Candidate genes were classified in two different ways; (i) phenotypic defects produced by a mutation on the mouse ortholog, or (ii) linkage of the human ortholog with a disease. A total of 4004 mouse orthologs with mutant phenotype data were collected from MGI (http://www.informatics.jax.org, see Materials and methods). We

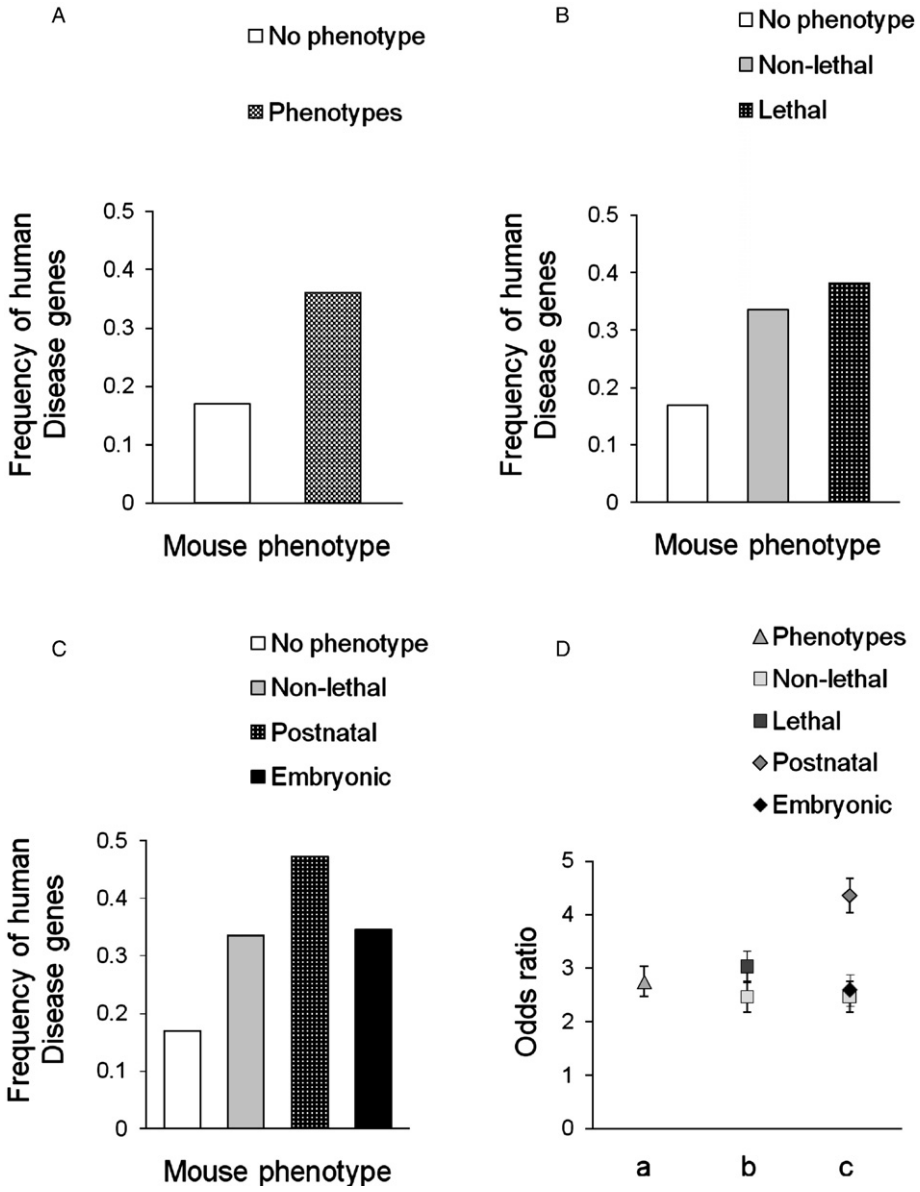

Fig. 1. The frequencies of human disease genes in different mouse *phenotype* groups were calculated and compared with those in *no phenotype* genes. *Phenotype* genes (A) were divided into *non-lethal* and *lethal* genes (B), which were further divided into *postnatal* and *embryonic lethal* genes (C). Statistical significance was tested by logistic regression model (D). On *X*-axis, 'a', 'b', and 'c' indicate different ways to divide *phenotype* genes corresponding to panels A, B, and C, respectively; *phenotype* (a), *non-lethal* and *lethal* (b), or *non-lethal*, *postnatal* and *embryonic lethal* genes (c). Confidence intervals are displayed with capped bars.

classified a gene as contributing to a discernible phenotype if a mutation on the mouse ortholog resulted in abnormal phenotype(s), including lethality. Hereafter, this group of genes will be abbreviated as 'P' genes. The genes at which null mutations in mouse do not result in any overt phenotypic abnormality were categorized as having *no phenotype* (NP). P genes were further divided into *lethal phenotype* (P-L) and *non-lethal phenotype* (P-NL) groups. In some subsequent analyses, P-L genes were further divided into *embryonic lethal* (P-L-Em) and *postnatal lethal* (P-L-Po) subgroups. Then, we determined whether those genes are linked with human diseases, using the list of 2789 disease genes obtained from HGMD (http://www.hgmd.cf.ac.uk/ac/index.php, see Materials and methods). All the genes and their information analyzed in this study were summarized in Supplementary Table 1A, 1B, 2A, and 2B (see Materials and methods).

Phenotype genes in mouse are more likely to be disease genes in human

We first investigated whether P genes (i.e. genes with fitness effects) in mouse are really more likely to be disease genes in human than NP genes (i.e. genes with no fitness effects). As shown in Fig. 1A, about 36% (1311/3635) of P genes versus only 17% (63/369) of NP genes were found to be human disease genes, indicating that P genes are significantly more likely to be disease genes ($p = 3.15 \times 10^{-6}$, odds ratio = 1.16, Fig. 1D). Statistical significance was estimated by logistic regression analysis (as described in Materials and methods). Next, we tested whether or not the severity of mouse phenotypes is associated with the frequency of disease genes. It was expected that the likelihood of a protein being involved in disease might scale with the probability of its gene to suffer mutation with large fitness effects [14]. Indeed, the frequencies of disease genes were about 38% (728/1904) and 34% (583/1731) in P-L and P-NL genes, respectively ($p = 3.86 \times 10^{-14}$, odds ratio = 3.02, $p = 8.40 \times 10^{-10}$, odds ratio = 2.47, respectively, Figs. 1B, D), suggesting that *lethal* genes, the genes related to more severe phenotypes in mouse, are more likely to be human disease genes.

The most essential genes, embryonic lethal genes, are less likely to be disease genes than postnatal lethal genes

Although *lethal*, so-called essential, genes contained more disease genes than P-NL genes, the difference between these groups was minute. To investigate this further, we compared two subgroups of P-L genes; P-L-Em and P-L-Po genes. Surprisingly, P-L-Po genes were found to have the highest frequency of disease genes (47%, 255/541), whereas P-L-Em genes contained only about 35% (473/1363) (Fig. 1C). The frequency of disease genes in P-L-Em genes was similar to that of P-NL genes (Fig. 1C), which explains the minute difference in the frequencies of disease genes between P-L and P-NL genes. These findings indicate

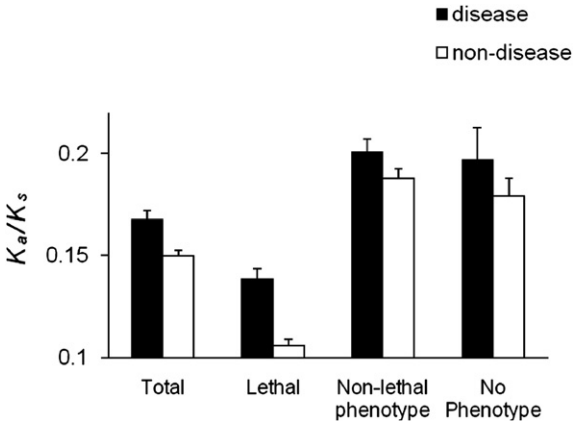

Fig. 2. K_a/K_s ratios were compared between disease and non-disease genes. The *X*-axis label indicates the mouse *phenotype* groups.

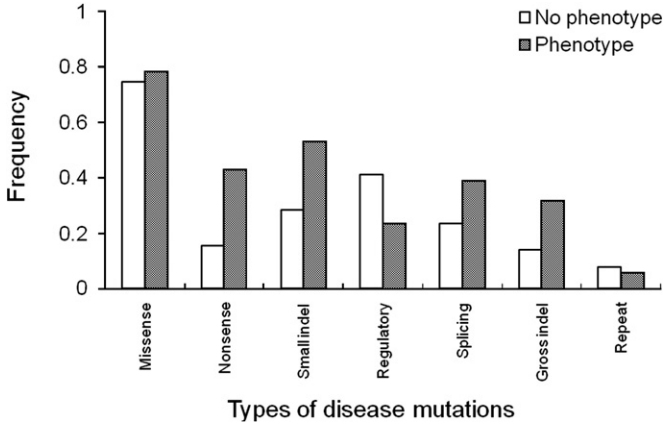

Fig. 3. The frequencies of various mutation types in disease genes were calculated and compared between *phenotype* and *no phenotype* genes.

that P-L-Em genes are less likely to be disease genes than P-L-Po genes. This inverse correlation between phenotypic severity in mouse and incidence of linkage with human disease may be due to strong negative selection (i.e., spontaneous abortions). Based on these observations, disease genes are apparently essential genes, given that P-L-Po genes, a part of *lethal* (i.e., essential) genes, showed the highest frequency of disease genes. Disease genes are, however, less likely to be extremely essential genes such as *embryonic lethal* genes.

Essential non-disease genes evolve more slowly than essential disease genes

Given that disease genes apparently display some characteristics of essential genes, we examined whether disease genes evolve more slowly than non-disease genes, similar to the manners reported for essential genes [3,10]. The evolutionary rate was calculated by K_a/K_s ratio using Li's method [18] between mouse and human orthologs, where K_a is the nonsynonymous substitution rate and K_s is the synonymous substitution rate (see Materials and methods). First, we confirmed that the genes with large fitness effect evolve more slowly, so that P-L-Em genes evolve the slowest ($K_a/K_s = 0.12$) with P-L-Po genes the next slowest ($K_a/K_s = 0.14$), and P-NL and NP genes having relatively high evolutionary rates ($K_a/K_s = 0.20$). As described previously [1,3,4], it was found that relatively nonessential genes (P-NL and NP genes) evolve rapidly compared to relatively essential genes (P-L-Em and P-L-Po genes). To examine whether disease genes evolve more slowly than non-disease genes, we compared K_a/K_s ratios between them (Fig. 2). Then, the comparison was repeated within each of phenotype subgroups. Nonessential genes did not show any statistically significant difference of K_a/K_s ratio between disease and non-disease genes (two-tailed *t* test, $p > 0.1$). However, non-disease genes in the P-L group were found to evolve more slowly than disease genes in the same group ($K_a/K_s = 0.106$ for non-disease versus $K_a/K_s = 0.136$ for disease, two-tailed *t* test, $p = 1 \times 10^{-9}$). Since non-disease genes evolve more slowly than disease genes in the P-L group (i.e. the group of essential genes), these essential non-disease genes seem to be under the influence of more stringent purifying selection than essential disease genes.

Types of disease mutations found in phenotype genes are different from those in no phenotype genes

To examine whether the types of disease mutations occurred in nonessential genes are different from those found in essential genes (i.e. genes with a fitness effect), we compared disease genes in P and NP groups. We calculated the frequency at which each type of disease mutations was found in disease genes of each group. As shown in Fig. 3, multiple types of mutations including nonsense, small indels, and

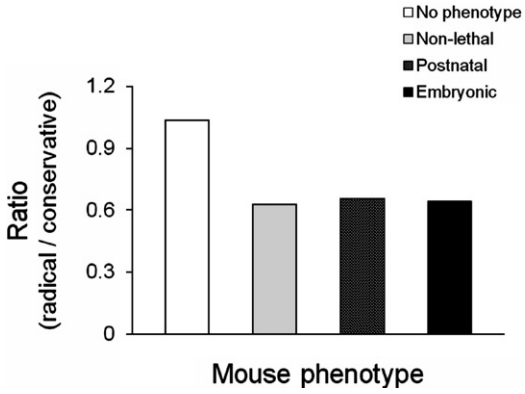

Fig. 4. The ratios of radical to conservative amino acid substitutions of disease mutations were obtained among different mouse phenotype groups. *Phenotype* genes were divided into *non-lethal*, *postnatal lethal* and *embryonic lethal* genes.

gross indels were found at significantly higher frequencies on the disease genes in the P group than those in the NP group. Interestingly, a significantly higher frequency of regulatory mutations was found on the disease genes in the NP group (Chi-squared test, $p<0.05$). Repeat mutations also showed higher frequency in NP genes, but significant conclusion could not be drawn largely due to the small sample size.

The disease genes in the no phenotype group have more radical amino acid mutations compared to those in the phenotype group

Fig. 3 indicates that the frequency of missense mutations on the disease genes in the P group did not significantly differ from those in the NP group (Chi-squared test, $p \gg 0.1$). Considering that NP genes encode relatively dispensable proteins and are subject to weaker purifying selection, it is reasonable to expect that these proteins require more radical amino acid changes to cause diseases. It is generally assumed that radical amino acid changes are much more likely to alter protein function than conservative substitutions [19–21]. To test this hypothesis, a total 21,807 amino acid changes nested in missense disease mutations were collected from HGMD, and were assigned into phenotype groups accordingly. We then calculated the ratio of radical amino acid substitutions to conservative amino acid substitutions for each group. The severity (radical *vs.* conservative) of amino acid changes was estimated based on Grantham's distance [20]. As hypothesized, missense disease mutations in NP genes were significantly more likely to be derived from radical amino acid changes (Fig. 4, Chi-squared test, $p \ll 0.0001$). There was no significant difference in the severity of amino acid changes among P-NL, P-L-Po, and P-L-Em groups (Fig. 4).

Discussion

Our results showed that P genes are more likely to be disease genes than NP genes, supporting the idea that human disease genes have properties of essential genes. Since P-L-Em genes were less likely to be disease genes than P-L-Po genes, our data support previous findings that disease genes are under an intermediate level of negative selection pressure [17, 22]. This idea is also supported by the fact that the difference of K_a/K_s ratios between disease and non-disease genes was dramatically significant in the P-L group, but not in both P-NL and NP groups. These non-disease genes seemed to be under the most stringent selection among all groups. Therefore, the lower K_a/K_s ratio of non-disease genes could be explained by the effects of extremely essential genes nested in non-disease genes outweighing those of relatively nonessential and dispensable genes in human. Taken together, all these findings consistently support the previous notion that disease genes have fitness effects, but may not be extremely essential especially during early development [8,17].

Genes contributing to organismic fitness are considered to be functionally important. A central tenet for the study of human gene function is that the orthologs of 'phenotype' genes of a model organism also perform an important analogous function [23]. As an estimation of the essentiality of a human gene, we used phenotypic consequences of mutations on mouse orthologs. Although the relatively short evolutionarily distance between human and mouse is an obvious advantage, it is clear some technical limits are placed on our analysis. First, the number of mouse genes with a known mutation phenotype, comprises less than 20% of the genome (4004 genes). Although this is relatively extensive, it may still have some bias towards obvious phenotype genes or human disease genes. Second, our analysis ignores certain aspects of the evolutionary process such as positive selections in human lineage. Third, our estimation on essentiality is not quantitative. Despite these methodological limitations, however, we believe this methodology represents a useful and complementary approach to existing methods.

In addition to the characteristics of human disease genes described above, the characteristics of mutations linked with human disease were compared between two subgroups of disease genes based on their essentiality; disease mutations in P and NP genes. The disease mutations in NP genes displayed significantly frequent regulatory mutations while those in P genes had a higher frequency of coding mutations in human disease genes. Since disease genes in the NP group are dispensable genes in mouse but may not be in human, this raises the possibility that such genes become important in human lineage by acquiring new, more important function(s) through positive selection. We hypothesize that these newly acquired functional changes may occur in regulatory regions. Indeed, it would be interesting to see what phenotypes would be revealed if regulatory regions in those genes were deleted or mutated in mice. Interestingly, missense mutations found in nonessential disease genes are more likely to be radical amino acid substitutions than those occurring in essential disease genes. This is consistent with the idea that the phenotypic severity of a genetic disease is determined by both the essentiality of the gene and the degree of deleterious effect of the mutation on the encoded protein. As previously noted [17], essentiality of disease genes might be better described on a continuous spectrum, although on average these genes remain intermediate to extremely essential and completely dispensable genes. Different mutations, as we describe here, may be one of the factors that broaden the spectrum of essentiality of disease genes.

Features of human disease genes have drawn much attention, since the signature found in known human disease genes can be used to identify novel genes associated with human disease. However, simple analysis of disease genes as a group lacks the power to accurately illustrate the complex scenarios associated with human disease genes. Along with previous studies using various classifications of human disease genes, we suggest classification of human disease genes based on its essentiality as a useful option for further studies. We believe the update of human disease genes combining with the development of the analysis tools will expand the understanding of evolutionary features in human disease genes in the future.

Materials and methods

Compiling lists of disease genes and mouse mutant phenotype genes

Lists of disease genes were obtained from HGMD (http://www.hgmd.cf.ac.uk/ac/index.php). A license for downloading the professional version of the database was obtained from the Biobase Co. and the databases were installed on a local sever computer. A total of 2789 genes were finally collected as human disease genes after removing genes with no sequence information. We also obtained various disease-associative information from the database such as types and their number of mutation found in each disease gene, gene symbol,

and gene length, etc. A total of 4004 of mouse phenotype related genes were obtained from MGI (http://www.informatics.jax. org/). We collected only genes with mutations derived from knock-out, gene-trap, and transgenic (gene-disruption) approaches, and classified them into different groups as described in the Results section. The mouse phenotype genes were linked to the human disease genes by gene symbols using a home-built perl script. Among all the studied genes, the human disease genes were listed in Supplementary Table 1 while the non-human disease genes were in Supplementary Table 2. Both Table 1A and 1B classify genes based on the severity of mouse phenotypes as described in the main text. Both tables contain gene name (HUGO gene symbol), RefSeq gene identifiers of mouse and human genes, mutation types and their frequencies found in human disease, and evolutionary rate values including K_a, K_s, K_a/K_s ratio. In addition, MGI and OMIM identifiers as well as gene descriptions were summarized in Supplementary Table 2A and 2B, for the human disease genes and the non-human disease genes, respectively.

Sequence collection and calculation of K_a/K_s ratio

To calculate K_a (nonsynonymous substitution rate) and K_s (synonymous substitution rate), we first obtained the coding sequences of orthologs between human and mouse. We only retrieved Refseq entries (ftp://ftp.ncbi.nih.gov) containing NM_ in the accession numbers. We referred 'gene_info' and 'gene2refseq' files downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov) to assign the gene symbols to the corresponding Refseq sequences. Then, the coding sequences were translated into amino acid sequences. Orthologs were identified based on reciprocal best hits by BLASTP search with a cutoff score of $E=10^{-20}$. The 4004 mouse sequences and their human orthologs were subtracted from the determined orthologs via perl script. Sequences were aligned through ClustalW, and the K_a/K_s ratio was calculated by the Li method [18].

Classification of amino acid replacement by Grantham's distance

A total of 21,807 missense mutations related to the human diseases were downloaded from the HGMD and classified according to Grantham's amino acid replacement distance matrix [19]. Following previous convention [20,21], a Grantham's distance of 100 or less was considered conservative and otherwise radical.

Statistical test

The logistic regression model was used to determine whether or not the phenotype genes in mouse are more likely to be disease genes in human. Let Y be a binary response variable representing whether a gene is linked to human disease or not. Explanatory variables represent four groups of essentiality of the genes in mouse: NP, P-NL, P-L-Po, and P-L-Em. Then, two types of logistic regression models were defined as follow:

$$\log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x, \quad (1)$$

$$\log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (2)$$

where $\pi(x) = P(Y=1|X=x) = 1-P(Y=0|X=x)$. In Model (1), one explanatory variable x, coded as 0 for NP, 1 for P-NL, 2 for P-L-Po, and 3 for P-L-Em, was used. Note that this model uses ordinal information in groups. On the other hand, Model (2) uses three indicator variables treating NP as a baseline group. That is, x_1 is 1 for P-NL and 0 otherwise; x_2 and x_3 are defined similarly for P-L-Po and P-L-Em, respectively.

From the estimators of βs, the estimators of the odds ratio (e^{β}) can be derived. When disease genes are more likely to be essential in mouse, the odds ratio would be larger than 1. The hypotheses for comparing groups are H_0: $\beta_1=0$ for Model (1) and H_0: $\beta_1=\beta_2=\beta_3=0$ for Model (2). We used statistical package R for the analysis (http://www.r-project.org/index.html).

Acknowledgments

This work was supported by the program of Nuclear R and D program (M2070600005-08B0600-00510) of the Korea Science and Engineering Foundation (KOSEF) and by the program of 2007 Research Grant from Kangwon National University. The work of Jungsun Park and Taesung Park was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126). We thank Noah M Walton and Tiffany M Carr for critically reading this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.08.001.

References

[1] I.K. Jordan, I.B. Rogozin, Y.I. Wolf, E.V. Koonin, Essential genes are more evolutionarily conserved than are nonessential genes in bacteria, Genome Res. 12 (2002) 962–968.
[2] S. Gerdes, R. Edwards, M. Kubal, M. Fonstein, R. Stevens, A. Osterman, Essential genes on metabolic maps, Curr. Opin. Biotechnol. 17 (2006) 448–456.
[3] A.E. Hirsh, H.B. Fraser, Protein dispensability and rate of evolution, Nature 411 (2001) 1046–1049.
[4] J. Yang, Z. Gu, W.H. Li, Rate of protein evolution versus fitness effect of gene deletion, Mol. Biol. Evol. 20 (2003) 772–774.
[5] C. Pal, B. Papp, L.D. Hurst, Genomic function: rate of evolution and gene dispensability, Nature 421 (2003) 496–497 discussion 497–498.
[6] E.P. Rocha, A. Danchin, An analysis of determinants of amino acids substitution rates in bacterial proteins, Mol. Biol. Evol. 21 (2004) 108–116.
[7] H.B. Fraser, D.P. Wall, A.E. Hirsh, A simple dependence between protein evolution rate and the number of protein–protein interactions, BMC Evol. Biol. 3 (2003) 11.
[8] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabasi, The human disease network, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8685–8690.
[9] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.
[10] D.P. Wall, A.E. Hirsh, H.B. Fraser, J. Kumm, G. Giaever, M.B. Eisen, M.W. Feldman, Functional genomic analysis of the rates of protein evolution, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 5483–5488.
[11] A.J. Butte, V.J. Dzau, S.B. Glueck, Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues", Physiol. Genomics 7 (2001) 95–96.
[12] J.A. Warrington, A. Nair, M. Mahadevappa, M. Tsyganskaya, Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes, Physiol. Genomics 2 (2000) 143–147.
[13] S. Bortoluzzi, C. Romualdi, A. Bisognin, G.A. Danieli, Disease genes and intracellular protein networks, Physiol. Genomics 15 (2003) 223–227.
[14] N. Lopez-Bigas, C.A. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic disease, Nucleic Acids Res. 32 (2004) 3108–3114.
[15] E.P. Rocha, The quest for the universals of protein evolution, Trends Genet. 22 (2006) 412–416.
[16] H. Huang, E.E. Winter, H. Wang, K.G. Weinstock, H. Xing, L. Goodstadt, P.D. Stenson, D.N. Cooper, D. Smith, M.M. Alba, C.P. Ponting, K. Fechtel, Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes, Genome Biol. 5 (2004) R47.
[17] Z. Tu, L. Wang, M. Xu, X. Zhou, T. Chen, F. Sun, Further understanding human disease genes by comparing with housekeeping genes and other genes, BMC Genomics 7 (2006) 31.
[18] W.H. Li, Unbiased estimation of the rates of synonymous and nonsynonymous substitution, J. Mol. Evol. 36 (1993) 96–99.
[19] T. Dagan, Y. Talmor, D. Graur, Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection, Mol. Biol. Evol. 19 (2002) 1022–1025.
[20] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.
[21] A.L. Hughes, J.A. Green, J.M. Garbayo, R.M. Roberts, Adaptive diversification within a large family of recently duplicated, placentally expressed genes, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 3319–3323.
[22] N.G. Smith, A. Eyre-Walker, Human disease genes: patterns and predictions, Gene 318 (2003) 169–175.
[23] A.E. Lathrop, L. Loeb, Further investigations on the origin of tumors in mice. III. On the part played by internal secretion in the spontaneous development of tumors, J. Cancer Res. 1 (1916) 1–19.